Understanding Social Movement by Keyword Tracking in Social media

Virach Sornlertlamvanich
virach@gmail.com
Kobkrit Viriyayudhakorn
kobkrit@gmail.com
THAILAND

67,091,000
TOTAL POPULATION

34%
URBAN

66%
RURAL

24,000,000
INTERNET USERS

36%
INTERNET PENETRATION

16,834,140
USERS ON TOP SOCIAL NETWORK

25%
SOCIAL NETWORKING PENETRATION

78,667,910
MOBILE SUBSCRIBERS

117%
MOBILE PENETRATION

http://tulaneict4d.wordpress.com/2013/04/05/social-media-in-thailand/

Linguistic and Cultural Diversity in Cyberspace, UNESCO Information for All Programme, Yakutsk, Russian, 28 June–3 July 2014
Monthly Active Users on PC in 2013

Linguistic and Cultural Diversity in Cyberspace, UNESCO Information for All Programme, Yakutsk, Russian, 28 June–3 July 2014
Thailand Monthly Active Users on Social Network and Messenger App in 2013


Linguistic and Cultural Diversity in Cyberspace, UNESCO Information for All Programme, Yakutsk, Russian, 28 June–3 July 2014
Data Preparation

- Word segmentation
- Keyword extracted from topic related documents (training set)
- Tweeter inquiry using the prepared topic related list of keywords
- Text similarity using GETA algorithm
GETA: Association Search Engine

- Proposed by Akihiko Takano (Takano, 2003)
- Back-and-forth searches between words and documents spaces
  - Set of keywords \( \rightarrow \) The top-n ranks of associated documents
  - Set of documents \( \rightarrow \) The top-m ranks of associated keywords
- Require a Word Article Matrix (WAM), and a similarity function
GETA: Association Search Engine

- Proposed by Akihiko Takano (Takano, 2003)
- Back-and-forth searches between words and documents spaces
  - Set of **keywords** → The top-n ranks of **associated documents**
  - Set of **documents** → The top-m ranks of **associated keywords**
- Require a **Word Article Matrix (WAM)**, and a **similarity function**
Word Article Matrix (WAM)

Building WAM is indexing

All Wikipedia Articles

All Pages’ Contents

Word seg. & Lemmatization
Thinking -> Think

All Words List

All Pages’ Titles

| Pages/Words | “Twitter” | “Tennis” | “Dollar” | “Google” | ...
|-------------|-----------|----------|----------|----------|...
| IT          | 2         | 0        | 1        | 4        |...
| Sport       | 0         | 2        | 1        | 0        |...
| Economics   | 0         | 0        | 2        | 0        |...
| ...         | ...       | ...      | ...      | ...      |...

A Wikipedia WAM
Linguistic and Cultural Diversity in Cyberspace, UNESCO Information for All Programme, Yakutsk, Russian, 28 June–3 July 2014
Similarity Functions

• Ranking weight in both words and documents spaces
• Equation form

\[
SIM(d, q) = \sum_{t \in q} \frac{wq(t, q) \cdot wd(t, d)}{\text{norm}(d, q)}
\]

• \(wd(t, d) = \) Weight of term \(t\) in document \(d\), \(wd(t, d) = 0\) if \(t \notin d\)
• \(wq(t, q) = \) Weight of term \(t\) in query sentence \(q\), \(wq(t, q) = 0\) if \(t \notin q\)
• \(\text{norm}(d, q) = \) Normalization function due to the different length of \(d\) and \(q\).
• Smart measure (Singhal et.al., 1996)

\[
\frac{1}{\text{avg}(f_d) + \theta(f_d - \text{avg}(f_d))} \sum_{t \in q \land t \in d} \log\left(\frac{N}{f_t}\right) \cdot \frac{1 + \log(f_{d,t})}{1 + \log(\text{avg}_{\omega \in d}(f_{d,\omega}))} \cdot \frac{1 + \log(f_{q,t})}{1 + \log(\text{avg}_{\omega \in q}(f_{q,\omega}))}
\]
• Dot Product (Wilkinson et.al., 1996)

\[
\sum_{t \in q \land t \in d} (w_{q,t} \cdot w_{d,t})
\]

Linguistic and Cultural Diversity in Cyberspace, UNESCO Information for All Programme, Yakutsk, Russian, 28 June–3 July 2014
Example GETA Calculation

SIM Function: Dot Product

\[ \sum_{t \in \text{Vec} \cap \text{Ed}} (w_{q,t} \cdot w_{d,t}) \]

A idea label
“Twitter has 800M dollars”

Word seg. & Lemmatization

\text{dollars} \rightarrow \text{dollar}

<table>
<thead>
<tr>
<th>Pages</th>
<th>Words</th>
<th>“Twitter”</th>
<th>“Tennis”</th>
<th>“Dollar”</th>
<th>“Google”</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td></td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Sport</td>
<td></td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Economics</td>
<td></td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

(n=2) Select Top-n
Most Associated Documents Ranking

<table>
<thead>
<tr>
<th>Pages</th>
<th>Dot Product Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>3</td>
</tr>
<tr>
<td>Sport</td>
<td>1</td>
</tr>
<tr>
<td>Economics</td>
<td></td>
</tr>
</tbody>
</table>

A Word Article Matrix (WAM)  
A similarity function  
Keywords  
Top-Ranked Associated Keywords  
Words space  
Documents space

Linguistic and Cultural Diversity in Cyberspace, UNESCO Information for All Programme, Yakutsk, Russian, 28 June–3 July 2014
Example GETA Calculation

SIM Function: Dot Product \[ \sum_{t \in q \cap d} (w_{q,t} \cdot w_{d,t}) \]

A idea label “Twitter has 800M dollars”

Word seg. & Lemmatization dollars -> dollar

<table>
<thead>
<tr>
<th>“Twitter”</th>
<th>“Tennis”</th>
<th>“Dollar”</th>
<th>“Google”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Pages\Words | “Twitter” | “Tennis” | “Dollar” | “Google” |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Sport</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Economics</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Wikipedia WAM

(V(n=2) Select Top-n Most Associated Documents Ranking)

<table>
<thead>
<tr>
<th>Pages</th>
<th>Dot Product Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>3</td>
</tr>
<tr>
<td>Sport</td>
<td>0</td>
</tr>
<tr>
<td>Economics</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Linguistic and Cultural Diversity in Cyberspace, UNESCO Information for All Programme, Yakutsk, Russian, 28 June–3 July 2014
Example GETA Calculation (Cont.)

SIM Function: Dot Product $\sum_{t \in q \land d} (w_{q,t} \cdot w_{d,t})$

Top-2 Most Associated Documents

<table>
<thead>
<tr>
<th>Pages</th>
<th>Dot Product Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>3</td>
</tr>
<tr>
<td>Sport</td>
<td>0</td>
</tr>
<tr>
<td>Economics</td>
<td>2</td>
</tr>
</tbody>
</table>

Output of GD

<table>
<thead>
<tr>
<th>Pages</th>
<th>“Twitter”</th>
<th>“Tennis”</th>
<th>“Dollar”</th>
<th>“Google”</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Sport</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Economics</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Wikipedia WAM

<table>
<thead>
<tr>
<th>“Twitter”</th>
<th>“Tennis”</th>
<th>“Dollar”</th>
<th>“Google”</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0</td>
<td>7</td>
<td>12</td>
</tr>
</tbody>
</table>

(m=3) Top-m Most Associated Keywords

Linguistic and Cultural Diversity in Cyberspace, UNESCO Information for All Programme, Yakutsk, Russian, 28 June–3 July 2014
**Example GETA Calculation (Cont.)**

SIM Function: Dot Product \( \sum_{t \in q \land d} (w_{q,t} \cdot w_{d,t}) \)

**Top-2 Most Associated Documents**

<table>
<thead>
<tr>
<th>Pages</th>
<th>Dot Product Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>3</td>
</tr>
<tr>
<td>Sport</td>
<td>0</td>
</tr>
<tr>
<td>Economics</td>
<td>2</td>
</tr>
</tbody>
</table>

*Sport Pages = 0, since only Top-2 is selected from previous slide.*

**A idea label**

"Twitter has 800M dollars"

1. Twitter has 800M dollars.

**Google, Dollar, Twitter**

**Wikipedia WAM**

*(m=3) Top-m Most Associated Keywords*
Tweeter Viewer

Word segmentation

Word

IT 2 0 1 4
Sport 0 2 1 0
Economics 0 0 2 0

Word Article Matrix (WAM)

Tweet query by keyword

Domain specific tweets

Social movement timeline viewer

IT 2014
Sport 2010
Economics 2020

Linguistic and Cultural Diversity in Cyberspace, UNESCO Information for All Programme, Yakutsk, Russian, 28 June–3 July 2014
Coup on May 22, 2014

- military, NCPO, country, announce, peace, power, government, coup d’etat, gathering, police, situation, PM, control, seize, meeting, economy, law, war, leader, minister, election, democracy, revolution, seize the power, curfew, martial law
Tweet Query

- Search Tweets by using Restful API
  - GET search/tweets. Set q = the keyword set
  - 100 tweets/search limited
  - Repeatedly fetch data until all tweets in the coup periods are discovered
- Be able to search back to 7 days

May 22, 2014 Coup-related tweet: 339,148 tweets
Linguistic and Cultural Diversity in Cyberspace, UNESCO Information for All Programme, Yakutsk, Russian, 28 June–3 July 2014
Conclusion

• Key word expansion is effective to understand the short message i.e. tweet
• Key word expansion can be done using the known training corpus and GETA algorithm
• Timeline word cloud shows the development of the social movement e.g. some events are predictable